HOW CAN IGNORANT BUT PATIENT COGNITIVE TERMINALS LEARN THEIR STRATEGY AND UTILITY?

S.M. Perlaza

France Telecom - Orange Labs
92794 Issy-les-Moulineaux, France
samir.medinaperlaza@orange-ftgroup.com

H. Tembine, and S. Lasaulce

CNRS – SUPELEC – Univ. Paris Sud
91190 Gif-sur-Yvette, France
{hamidou,lasaulce}@lss.supelec.fr

ABSTRACT

This paper aims to contribute to bridge the gap between existing theoretical results in distributed radio resource allocation policies based on equilibria in games (assuming complete information and rational players) and practical design of signal processing algorithms for self-configuring wireless networks. For this purpose, the framework of learning theory in games is exploited. Here, a new learning algorithm based on mild information assumptions at the transmitters is presented. This algorithm possesses attractive convergence properties not available for standard reinforcement learning algorithms and in addition, it allows each transmitter to learn both its optimal strategy and the values of its expected utility for all its actions. A detailed convergence analysis is conducted. In particular, a framework for studying heterogeneous wireless networks where transmitters do not learn at the same rate is provided. The proposed algorithm, which can be applied to any wireless network verifying the information assumptions stated, is applied to the case of multiple access channels in order to provide some numerical results.

1. INTRODUCTION

Concepts such as cognitive radio, unlicensed bands, ad hoc networks, and self-configuring networks are becoming more and more important in the wireless communications arena. One common point between the corresponding scenarios is that radio devices autonomously set up their transmission configuration and interact with each other through mutual interference. Hence, game theory, a branch of mathematics analyzing interactions between (inter-dependent) decision makers, appears as a natural paradigm to analyze, optimize and design these types of scenarios. This is one of the reasons why it is applied more and more intensively to wireless networks (see e.g., [1] and references therein). For instance, many distributed power control and resource allocation schemes have been proposed by exploiting concepts of game theory such as equilibria. Nonetheless, there is a quite general consensus to say that the corresponding analysis are useful but rely on strong information and behavior assumptions on the terminals and thus, predicted game outcomes (most often Nash equilibria) will be rarely observed. Fortunately, specialized sub-branches of game theory, such as algorithmic game theory [2], learning theory in games [3], and mechanism design [4] have been developed before researchers and engineers encounter this problem. One of the many purposes of learning theory is to design algorithms exploiting partial information on the game to converge to solutions predicted under full information assumption. Note that learning theory is still a theory in the sense that the designed algorithms generally require a large number of steps to converge and therefore needs to be improved to be fruitfully implemented. One of our objectives in this paper is precisely to propose and analyze a new learning algorithm which relies on mild information assumptions from the standpoint of a wireless terminal designer.

In the available literature on wireless communications, the closest work to the one reported here is [5]. The authors of [5] proposed to apply the reinforcement learning algorithm (RLA) initially introduced by [6] and revisited by [7] in order to propose a distributed (always in the sense of the decision) power control policy for networks with finite numbers of transmitters and power levels. Based on the sole knowledge of the value of its utility at each step, the corresponding algorithm consists, for each transmitter, in updating its strategy, namely its probability distribution over the possible power levels. The authors of [5] conduct the convergence analysis for $2-$player $2-$action games. In the present paper we propose a new learning algorithm based on the idea of $Q-$learning used in Markov decision processes [8] and the Boltzmann-Gibbs learning algorithm [9]. The main reasons for proposing this new algorithm, which is detailed in Sec. 3.1, are the following: (1) The proposed algorithm converges for a class of games broader than the one of the RLA; (2) It allows a transmitter to learn its distributed strategy but also its expected utility function for all its actions. In addition to this algorithm, we provide several general convergence results. In particular, we give a theoretical framework for analyzing networks comprising transmitters who learn at different rates.

2. PRELIMINARY PART

The purpose of this section is twofold. The first two subsections are dedicated to readers who are not familiar with game theory and reinforcement learning. Next, we describe the assumptions required for our algorithm to be applied to a given wireless network. We do not provide a specific signal model at this point, since our algorithm can be applied to many types of self-configuring wireless networks.

2.1. Review of basic game-theoretic concepts

The three basic components of a game are the set of players ($\mathcal{K} = \{1, ..., K\}$, K being the number of players), the action spaces ($\mathcal{A}_1, ..., \mathcal{A}_K$), and the cost/payoff/reward/utility functions ($u_1, ..., u_K$); in this paper only discrete action spaces are considered and are denoted by $\mathcal{A}_i = \{A_{i,1}, ..., A_{i,|\mathcal{A}_i|}\}$

for player $i \in \mathcal{K}$. This paper covers scenarios where players are transmitters who are able to choose their actions by themselves. The actions can be typically a power level, a vector of powers, the constellation size, or any other transmission parameter. The utility function can be, for example, the transmission rate (see e.g., [10]) or energy-efficiency (see e.g., [11]). If the game is played only once, the game is called a static or one-shot game and can be represented by a 3-tuple in its strategic-form: $\mathcal{G} = (\mathcal{K}, \{\mathcal{A}_i\}_{i\in\mathcal{K}}, \{u_i\}_{i\in\mathcal{K}})$. When player $i \in \mathcal{K}$ chooses an action in $\mathcal{A}_i$ according to a probability distribution

$$\underline{\pi}_i = \left(\pi_{i,1}, ..., \pi_{i,|\mathcal{A}_i|}\right) \quad (1)$$

over $\mathcal{A}_i$, the choice of $\underline{\pi}_i$ is called a mixed strategy. When $\underline{\pi}_i$ is on a vertex of the simplex $\Delta(\mathcal{A}_i)$ the mixed strategy boils down to a pure strategy i.e., the deterministic choice of an action. In this paper we consider dynamic games i.e., players play several times, and at each time t each player $i \in \mathcal{K}$ chooses its action $a_i(t) \in \mathcal{A}_i$ following its probability distribution $\underline{\pi}_i(t)$. Such a choice can be made every symbol, block of symbols, or frame duration. Additionally, the dynamic game is stochastic in the sense that the game can have a state which changes from time to time. Often, in wireless communications, the state will be typically the channel state e.g., the vector of channel gains $\underline{h}(t) = (h_1(t), ..., h_K(t))$. This is why we will denote the instantaneous utility function by $u_i^{\underline{h}}$. Neither the overall channel state $\underline{h}$ nor the individual channel state h_i is known to transmitter $i \in \mathcal{K}$. The channel transition probability is also assumed to be unknown to every transmitter. To conclude this section we would like to mention that often, it is assumed that players are rational and that rationality is common knowledge. In this paper, each transmitter does the best for himself but does not need to know whether the other transmitters are rational or not.

2.2. Review of the reinforcement learning algorithm of [7]

The utility function at a given time t depends on the actions played by the different players. By denoting $a_i(t)$ the action played by i at time t, we can write the utility of player i as $u_i^{\underline{h}(t)}(a_1(t), ..., a_K(t))$. By notational abuse but for the sake of clarity the utility function of i at stage t will be denoted by $u_i(t)$. For example, with our notations we have that $\pi_{i,1}(t) = \Pr[a_i(t) = A_{i,1}]$. The RLA of [6, 7, 5] consists in updating the probability distribution over the possible actions as follows: $\forall i \in \mathcal{K}, \forall j \in \{1, ..., |\mathcal{A}_i|\}$,

$$\pi_{i,j}(t+1) = \pi_{i,j}(t) + \lambda_i(t) u_i(t) \left(\mathbb{1}_{a_i(t)=A_{i,j}} - \pi_{i,j}(t)\right) \quad (2)$$

where $\mathbb{1}$ is the indicator function and $0 < \lambda_{i,t} < 1$ is the weight chosen by player i at stage t; this parameter has to be normalized and represents the learning rate. The algorithm is simple to interpret: the action which was played at the last stage, namely $a_i(t)$, sees its probability increased (since $\left(\mathbb{1}_{a_i(t)=A_{i,j}} - \pi_{i,j}(t)\right) \geq 0$) while the other actions see their probability decreased (since $0 - \pi_{i,j}(t) \leq 0$). The key point here is that the increment in the probability of each action $A_{i,j}$ depends on the corresponding observed utility and its learning rate. More importantly, note that in (2), for each player, only the value of its individual utility function at stage t is required. Therefore, the knowledge of the utility function u_i is not assumed for implementing the algorithm. This is one of the reasons why gradient-like techniques are not applicable here.

2.3. Scope of the paper

At this point, it is possible to delineate the framework of this paper. The results provided here apply to all wireless scenarios meeting the following conditions:

- The addressed wireless game must be **finite** i.e., both the number of transmitters and possible actions must be finite. For example, this is a suited assumption if the action is a transmit power level, a modulation constellation size, a number of transmit antennas to use, or a number of receive base stations to be connected to. However the number of channel states $\underline{h}$ can be arbitrary (finite or infinite).
- Each transmitter must be able to observe the **value** of its individual utility obtained at each time t. A very practical example is the frame success rate. If the transmitter is acknowledged by the receiver frame by frame by an ACK/NACK (acknowledgment / non-acknowledgment) message, then the transmitter is therefore able to know the instantaneous value of the number of successfully received frames.
- The **duration of the interaction** between the transmitters must be sufficiently high in order to observe the convergence of the strategies (probability distributions $\pi_k(t)$, for all $k \in \mathcal{K}$) and therefore achieve an equilibrium. The authors insist on the fact that one of the purposes of this paper is to promote learning theory in the wireless community as an intermediate theory to bridge the gap between equilibrium-based (e.g., control or resource allocation) distributed policies and implementable signal processing algorithms and propose and analyze a new algorithms bridging this gap.

3. JOINT UTILITY STRATEGY ESTIMATION BASED REINFORCEMENT LEARNING

3.1. The algorithm

As mentioned in the previous section we consider a K-player stochastic dynamic game with an arbitrary number of channel states. The players are the transmitters and the utility function is the average utility of transmitter i, that is, $\frac{1}{T}\sum_{t=1}^{T} u_i(t)$. Based on the sole knowledge of the value of $u_i(t)$, each transmitter updates its mixed strategy at stage $t+1$. As mentioned in Sec. 2.2 the RLA of [6, 7, 5] does not converge in all (wireless) games. Our objective is to design an algorithm which converges for a broader class of games and learns/estimates not only the strategies of the transmitter but also its expected utility for all its actions. Indeed, for existing RLA only the strategy is learned (reinforced). Here, we propose to reinforce both strategy and utility, which is why we call the proposed algorithm Joint Utility STrategy Estimation (JUSTE) based RLA. First, we provide the new learning algorithm and then we explain how it has been built. For each $i \in \mathcal{K}$, $j \in \{1, ..., |\mathcal{A}_i|\}$ the probability transmitter i assigns to action j is updated according to:

$$\left\{\begin{array}{lcl} \pi_{i,j}(t+1) & = & [1-\lambda_i(t)]\pi_{i,j}(t) + \lambda_i(t)\beta_i\left(\widehat{u}_{i,j}(t)\right) \\ \widehat{u}_{i,j}(t+1) & = & \frac{\mu_i(t)}{\pi_{i,j}(t)}\mathbb{1}_{a_i(t)=A_{i,j}}\left[u_{i,j}(t) - \widehat{u}_{i,j}(t)\right] \\ & & +\widehat{u}_{i,j}(t) \end{array}\right. \quad (3)$$

where $\lambda_i(t)$ and $\mu_i(t)$ are respectively the learning rates of the strategy and utility of transmitter $i \in \mathcal{K}$ at time t. $\widehat{u}_{i,j}(t)$ is the estimated utility of transmitter $i \in \mathcal{K}$ for action $A_{i,j}$ at time t. β_i is the Boltzmann-Gibbs (BG) distribution. The BG distribution associated with a vector $\underline{x} = (x_1, ..., x_M)$ is given by $\forall m \in \{1, ..., M\}, \beta(x_m) = \frac{e^{\alpha x_m}}{\sum_{n\in\{1,...,M\}} e^{\alpha x_n}}$ where α is a parameter (α represents the temperature in physics, rationality level in learning, etc). In our case we have that

$$\beta_i(\widehat{u}_{i,j}(t)) = \frac{e^{\alpha \widehat{u}_{i,j}(t)}}{\sum_{k\in\{1,...,|\mathcal{A}_i|\}} e^{\alpha \widehat{u}_{i,k}(t)}}. \quad (4)$$

The proposed algorithm, which can be initialized, in an arbitrary way is inspired from $Q-$learning algorithms used in Markov decision processes to estimate a function [8] (here $u_{i,j}$) and the BG learning used in games to learn strategies only (which can be recovered by choosing $\lambda_{i,j} \to 1$ and ignoring the second equation). To implement the proposed algorithm only the knowledge of the value of the individual utility associated with the played action, namely $a_i(t)$ at stage t, is required. An important difference between JUSTE-based RLA and conventional RLA is that the values of $u_{i,j}$ for the actions not played at stage t, namely $A_{i,j} \neq a_i(t)$, are now estimated. The corresponding estimates are used to regulate the convergence process of the strategy. Indeed, these estimates allow the probability associated with a given action to be updated more frequently (not only when the associated action is drawn). The choice consisting in coupling $Q-$learning and BG learning is quite subtle. Indeed, with BG learning, probabilities never vanish and can therefore used at the denominator in the second equation of (3). To conclude this section we would like to add a comment on learning rates $\lambda_{i,j}(t)$, $\mu_{i,j}(t)$. As the algorithm updates probabilities, normalized learning rates have to be chosen. Additionally, if one wants to guarantee the algorithm to converge to a solution of a differential equation, the following conditions have to be met for $\lambda_{i,j}(t)$:

$$\sum_{t\geq 0} \lambda_{i,j}(t) = +\infty, \ \sum_{t\geq 0} \lambda_{i,j}^2(t) < +\infty. \quad (5)$$

The same conditions are required for $\mu_{i,j}(t)$. A simple example is $\lambda_{i,j}(t) = \frac{1}{(t+c)^{\gamma_i}}$ with $c > 0$ and $\frac{1}{2} < \gamma_i \leq 1$.

Remark: We use the BG instead of standard RLA because when considering the standard RLA for both utility and strategies, one gets

$$\left\{ \begin{array}{lcl} \pi_{i,j}(t+1) & = & \lambda_i(t) u_i(t) \left[\mathbb{1}_{a_i(t)=A_{i,j}} - \pi_{i,j}(t)\right] \\ & & +\pi_{i,j}(t) \\ \widehat{u}_{i,j}(t+1) & = & \mu_i(t) \mathbb{1}_{a_i(t)=A_{i,j}} \left[u_i(t) - \widehat{u}_{i,j}(t)\right] \\ & & +\widehat{u}_{i,j}(t) \end{array} \right. \quad (6)$$

which leads to a composition of re-scaled replicator dynamics. However, the replicator dynamics may not lead to equilibria (for example the faces of the simplex are forward invariant). Using a smooth mapping like BG, we deviate the trajectory to the relative interior of the simplex. Thus, any rest point of our algorithm leads to a $\frac{1}{\alpha}-$equilibrium which gives an equilibrium when α goes to infinity.

3.2. Convergence analysis

Ensuring the convergence in a sufficiently broad class of games in wireless communications and ensuring the latter to be relatively fast w.r.t. existing solutions are fundamental issues. As mentioned in Sec. 2, we consider the general case where channel states are time-variant and thus, we focus on the convergence of each individual utility vector $\underline{u}_i = (u_{i,1}, ..., u_{i,|\mathcal{A}_i|})$ to $\mathbb{E}_{\underline{h}}\left[\underline{u}_i^{\underline{h}}\right]$. In the following, we provide several important results whose proofs are not provided because of the lack of space but some key elements are given.

Proposition 1 (Consequences of convergence) *If the JUSTE based RLA converges, then: (i)* $\lim_{t\to+\infty} \widehat{\underline{u}}_i(t) = \mathbb{E}_{\underline{h}}[\underline{u}_i^{\underline{h}}]$*; (ii) the limit strategy correspond to a BG equilibrium that is the unique solution of the fixed point equation* $\beta_i(\mathbb{E}_{\underline{h},\underline{\pi}}[u_{i,j}^{\underline{h}}]) = \pi_{i,j}$ *with* $\underline{\pi} = (\pi_1, ..., \pi_K)$.

This result can be proved by using results from the theory of stochastic approximation and more precisely the development of the ordinary differential equation (ODE) approach to stochastic approximation provided in [12, 13]. The main idea is to rewrite the second equation of (3) in Robin-Monro iterative form (see [13]) and approximate it by an ODE. In this paper, the two points we want to emphasize are the following: the proposed estimation procedure is consistent for the utilities; every rest point of the ODE is an equilibrium (this property is not verified by standard RLA which are approximated by the replicator dynamics [14].

Proposition 2 (Convergence to ODE) *Let* $\lambda_i(t) = k_i\lambda(t)$ *and* $\mu_i(t) = \ell_i\mu(t)$ *verify (5). Then the JUSTE based RLA converges almost surely and the limit utility and strategies are the solutions of the following system of ODEs:*

$$\left\{ \begin{array}{lcl} \frac{\mathrm{d}\pi_{i,j}(t)}{\mathrm{d}t} & = & k_i[\beta_i(\widehat{u}_{i,j}(t)) - \pi_{i,j}(t)] \\ \frac{\mathrm{d}\widehat{u}_{i,j}(t)}{\mathrm{d}t} & = & \ell_i[\mathbb{E}_{\underline{h},\underline{\pi}}[u_{i,j}^{\underline{h}}] - \widehat{u}_{i,j}(t)]. \end{array} \right. \quad (7)$$

In particular, $\lambda_i(t) = \mu_i(t)$ *implies that:*

$$\left\{ \begin{array}{lcl} \frac{\mathrm{d}\pi_{i,j}(t)}{\mathrm{d}t} & = & \beta_i(\widehat{u}_{i,j}(t)) - \pi_{i,j}(t) \\ \frac{\mathrm{d}\widehat{u}_{i,j}(t)}{\mathrm{d}t} & = & \mathbb{E}_{\underline{h},\underline{\pi}}[u_{i,j}^{\underline{h}}] - \widehat{u}_{i,j}(t). \end{array} \right. \quad (8)$$

Proposition 3 (Slow and fast learners) *Let* $\lambda_i(t)$ *and* $\mu_i(t)$ *verify (5). Additionally assume that* $\mu_i(t)$ *are equal to* $\mu(t)$ *and* $\lim_{t\to+\infty} \frac{\lambda_i(t)}{\lambda_{i+1}(t)} = 0$*. Then the JUSTE based RLA converges almost surely and the limit strategies are the solutions of the following ODE:*

$$\frac{\mathrm{d}\pi_{1,j}(t)}{\mathrm{d}t} = \beta_1 \circ \beta_2 \circ \ldots \circ \beta_K \left(\mathbb{E}_{\underline{h},\pi_1}[u_{1,j}^{\underline{h}}]\right) \quad (9)$$

In particular for two-player case, one has the system

$$\frac{\mathrm{d}\pi_{1,j}(t)}{\mathrm{d}t} = \beta_1 \left(\beta_2(\mathbb{E}_{\underline{h},\pi_1}[u_{1,j}^{\underline{h}}])\right) \quad (10)$$

The proof is built on the arguments given in [15, 13]. We see that for a sufficiently high number of stages, the optimal strategy of a transmitter is the solution to an ODE based on

the composition of the functions β_i. The proposed approach therefore gives a general framework for analyzing heterogeneous wireless networks where transmitters do not learn at the same rate. In the case where there exists a hierarchy in terms of learning rate between the transmitters, as defined in the above proposition, a very elegant solution can be obtained.

Proposition 4 (Sufficient conditions for conv. to equilibria) *If one of the following conditions is satisfied then the JUSTE based RLA converges:*
(i) the game is a dominance solvable game;
(ii) the game is potential;
(iii) the game is a two-transmitter zero-sum game;
(iv) the transmitters have only two actions and the network is composed of two types of learners (slow and fast ones);
(v) any finite game with unique evolutionarily stable strategy.

The class of games (i) corresponds to games where the iterated dominance procedure leads to a unique prediction of the game outcome: the main idea is that dominated strategies (a strategy is dominated for a given transmitter if there is another one that is better whatever the other transmitters do) can be removed. The power allocation game of [10] for multiple input multiple output channels is an example of this type of games (but not finite). Class (ii): recall that a game defined by $\left(\mathcal{K}, (\mathcal{A}_i)_{i\in\mathcal{K}}, (u_i)_{i\in\mathcal{K}}\right)$ is an exact potential game if there exists a function ϕ such that $\forall i \in \mathcal{K}$, $\forall \underline{a} \in \mathcal{A}_1 \times ...\mathcal{A}_K$, $\forall a_i' \in \mathcal{A}_i$, $u_i(\underline{a}) - u_i(a_i', \underline{a}_{-i}) = \phi(\underline{a}) - \phi(a_i', \underline{a}_{-i})$. The power control game of [16] and the wireless routing game of [17] are examples of potential games. Classes (iii) and (iv) do not call for particular comments. Class (v) corresponds to games with unique equilibrium which is resilient to small perturbation.

Proposition 5 (Convergence time) *The convergence time of the ODE of the strategy and utility for JUSTE based RLA to be $\epsilon-$close to the solutions are respectively given by:*
(i) $\Theta(\log(\frac{1}{\epsilon^2}))$; *(ii)* $\Theta(\log(\frac{1}{\epsilon}))$.

This proposition quantifies the intuition that probability distributions (strategies) need more time than utilities to be learned. Interestingly, the fact that JUSTE based RLA learn both strategies and utilities does not slow down the convergence process w.r.t. standard RLA, since (ii) correspond to the convergence time of the latter. To conclude this section let us mention a simple example in which the proposed algorithm outperform existing ones.

Example. Consider the power control game with 2 transmitters and 3 possible power levels, one can show that by slightly modifying the utility as $u_i(a_1, ..., a_K) = \mathbb{1}_{\{\mathrm{SINR}_i(\underline{a})>\gamma_0\}} + \mathbb{1}_{\{a_i<\max_{i'\neq i} a_{i'}\}} - \mathbb{1}_{\{\min_{i'\neq i}\mathrm{SINR}_{i'}(\underline{a})<\gamma_0\}}$, the game is not potential anymore and conventional learning algorithms such as standard RLA, replicator dynamics, log-linear dynamics, logit dynamics, best-response dynamics, and fictitious play fail to converge while the JUSTE based RLA converges; γ_0 is a target SINR, a_i represents the transmit power of i.

4. EXAMPLE: THE PARALLEL MULTIPLE ACCESS CHANNELS

Consider a set $\mathcal{K} = \{1, \ldots, K\}$ of transmitters sending independent signals to a set $\mathcal{S} = \{1, \ldots, S\}$ of receivers (e.g., access-points in Wi-Fi networks or base stations in cellular networks), each one operating in a dedicated frequency band and connected to the same network (through a radio network controller). For all $(i, s) \in \mathcal{K} \times \mathcal{S}$, denote by $h_{i,s}(t)$ the channel gain for transmitter i over receiver s at instant t. Let also $\underline{p}_i(t) = (p_{i,1}(t), \ldots, p_{i,S}(t))$ be the transmit power vector of player i. Here, $p_{i,s}(t) \geqslant 0$ represents the transmit power of player i over channel s at instant t and $\sum_{s\in\mathcal{S}} p_{i,s}(t) \leqslant p_{i,\max}$. Based on the discussion in [17], players are restricted to transmit at full power over a unique channel aiming to maximize its own transmission rate (utility function) $u_i(t) = \sum_{s\in\mathcal{S}} \log\left(1 + \frac{|h_{i,s}|^2 p_{i,s}}{\sigma_s^2 + \sum_{k\in\mathcal{K}\setminus\{i\}} |h_{k,s}|^2 p_{k,s}}\right)$, where $\underline{h} = (\underline{h}_1 ... \underline{h}_K)$ and $\underline{h}_i = (h_{i,1} ... h_{i,S})$. We write the strategy set of player $i \in \mathcal{K}$ as, $\mathcal{A}_i = \{p_{\max}\, \boldsymbol{e}_s : \forall s \in \mathcal{S},\ \boldsymbol{e}_s = (e_{s,n})_{n\in\mathcal{S}}$ and $\forall r \in \mathcal{S} \setminus s, e_{s,r} = 0$, and $e_{s,s} = 1\}$.

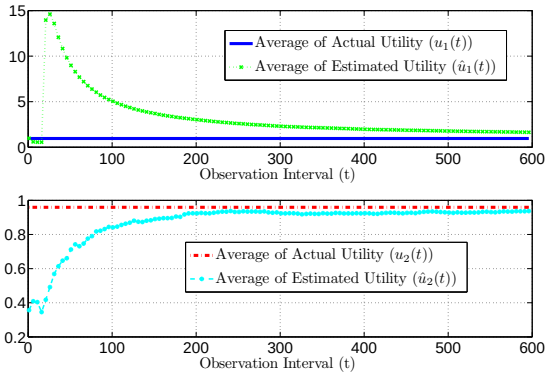

Fig. 1. Estimated and actual values of the average individual utilities as a function of time (measured in utility observation intervals), when $K = S = 2$, average signal to noise ratio $\mathrm{SNR}_i = \frac{p_{i,\max}}{\sigma^2} = 10$ dB.

In [17], it has been shown that the game $\mathcal{G} = (\mathcal{K}, \{\mathcal{A}_i\}_{i\in\mathcal{K}}, \{u_i\}_{i\in\mathcal{K}})$ is an exact potential game with multiple NE. Thus, the JUSTE-based RLA converges to one of the NE (Prop. 4). In the following, assume that all transmitters $i \in \mathcal{K}$ play the game by using our JUSTE-based RLA to update their probability distributions $\underline{\pi}_i(t)$.

For the sake of simplicity, we present some convergence results considering only $K = 2$ transmitters (players) and $S = 2$ receivers. In Fig. 1, the estimated and actual individual average utilities are plotted for both transmitters. The former is calculated based on the estimation of the utility of each player $i \in \mathcal{K}$ and the corresponding probability distribution $\underline{\pi}_i(t)$. The latter is the mean of all the utility observations during the whole transmission duration. In Fig. 1, we observe how the estimated average of the individual utilities (and thus, the strategies of all players) converge to the actual average individual utilities. One can say that convergence is relatively quick considering that each player only possesses information on its own strategy set and sporadic numerical observations of its utility.

In order to give an idea of the convergence time, we adopt the following assumption: once the error between the estimated and actual individual average utilities of all players is smaller than 5%, then it can be said that the network converged to an equilibrium configuration (ϵ-equilibrium). In

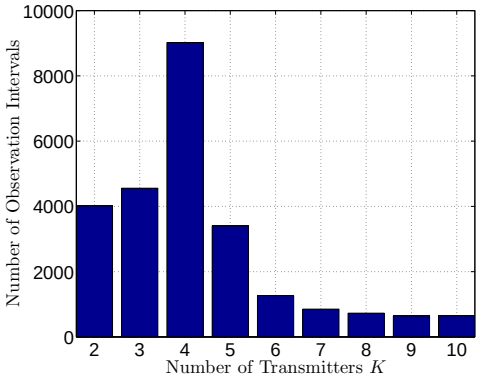

Fig. 2. Convergence time (measured in utility observation intervals) as a function of the number of transmitters in the network $K = \{2, \ldots, 10\}$ when $S = 3$ and average signal to noise ratio $\mathrm{SNR}_i = \frac{p_{i,\max}}{\sigma^2} = 10\mathrm{dB}$.

Fig. 2, we plot the convergence time as a function of the number of transmitters $K \in \{2, \ldots, 10\}$, when the number of receivers is kept constant $S = 3$. Therein, it can be shown that for weakly loaded networks, $\frac{K}{S} < 1$, the convergence time is larger. This is due to the fact that for weakly loaded networks transmitters are tempted to constantly change the receiver since there always exist a receiver with constantly time-varying unused channels.

5. CONCLUSION

As a standard RLA, the proposed learning algorithm (JUSTE based RLAs) only requires the knowledge of the value of the individual obtained utility associated with the latest action, without assuming rationality is common knowledge. The proposed algorithm is shown to have attractive convergence properties and to be suited to analyze complex scenarios such as those where transmitters do not learn at the same rate. A simple example was provided to show that conventional learning algorithms fail to converge in scenarios where JUSTE based RLAs do. The proposed framework seems to be very fruitful for designing implementable distributed control and allocation algorithms. One important technical challenge remains to accelerate the convergence rate of learning algorithms and design algorithms offering desired trade-offs between knowledge at the transmitters and convergence rate. In a general manner, game theory and learning theory in games seem to open a large avenue for technical innovation in terms of distributed control and radio resource allocation.

6. REFERENCES

[1] S. Lasaulce, M. Debbah, and E. Altman, "Methodologies for analyzing equilibria in wireless games," *IEEE Signal Processing Magazine, Special issue on Game Theory for Signal Processing*, Sep. 2009.

[2] Noam Nisam, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, *Algorithmic Game Theory*, Cambridge University Press, September 2007.

[3] Drew Fudenberg and David K. Levine, *The Theory of Learning in Games*, MIT Press, 1998.

[4] E. Maskin, "Nash equilibrium and mechanism design," *MIMEO, Institute for Advanced Study, School of Social Science*, 2008.

[5] Yiping Xing and R. Chandramouli, "Stochastic learning solution for distributed discrete power control game in wireless data networks," *IEEE/ACM Trans. Networking*, vol. 16, no. 4, pp. 932–944, 2008.

[6] Mosteller F. Bush R., "Stochastic models of learning.," *Wiley Sons, New York.*, 1955.

[7] P.S. Sastry, V.V. Phansalkar, and M.A.L. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 5, pp. 769–777, May 1994.

[8] Richard S. Sutton and Andrew G. Barto, "Reinforcement learning: An introduction," *MIT Press, Cambridge, MA*, 1998.

[9] Gibbs J. Willard, "On the equilibrium of heterogeneous substances," *Connecticut Acad. Sci.*, 1875-1878.

[10] E. V. Belmega, S. Lasaulce, and M. Debbah, "Power allocation games for MIMO multiple access channels with coordination," *IEEE Trans. on Wireless Communications*, vol. 6, pp. 3182–3192.

[11] D. Goodman and N. Mandayam, "Power control for wireless data," *IEEE Personal Communications*, p. 4854, April 2000.

[12] H. J. Kushner and D. S. Clark, "Stochastic approximation methods for constrained and unconstrained systems," *Springer, New York*, 1978.

[13] M. Benaïm, "Dynamics of stochastic approximations.," *Le Seminaire de Probabilites. Lectures Notes in Mathematics*, vol. 1709, 1999.

[14] Taylor and Jonker, "Evolutionarily stable strategies and game dynamics," *Mathematical Bioscience*, vol. 40, pp. 145–156, 1978.

[15] Borkar V. S., "Stochastic approximation with two timescales," *Systems Control Lett.*, vol. 29, pp. 291–294, 1997.

[16] G. Scutari, S. Barbarossa, and D.P. Palomar, "Potential games: A framework for vector power control problems with coupled constraints," *in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, May 2006.

[17] S. M. Perlaza, E. V. Belmega, S. Lasaulce, and M. Debbah, "On the base station selection and base station sharing in self-configuring networks," *3rd ICST/ACM International Workshop on Game Theory in Communication Networks*, Oct. 2009.